Loan Risk Identification

Identification of risk factors involved with default loans &
The predicting models that could help investors avoid them

Executive Summary For LendingClub Investors

By Eran Schenker

Project Overview

- Background & Business challenge
- Data Exploration & Preliminary Analysis
- Baseline Modeling
- Feature Engineering
- Building a Predictive Model
- Model Performance
- Summary and Final Recommendations
- Next Steps
- Supplementary Slides

Background & Business Challenge

- Lending Club (LC)
  I. The largest online loan platform ($10.9 B new loans -2018)
  II. Peer-to-peer (P2P) lending site, directly connects borrowers and investors

- Value Proposition & Challenge:
  I. “Cutting out the middleman” → opens possibilities for borrowers and investors
  II. Exposes investors to the risk of unpaid loans → i.e. **Default Loans** (6% of all loans)

- Project outline:
  I. Built a classification model based on data from LC website
  II. Projects whether a **loan** in 2014 → **Fully paid / Default loan**
  III. analysis used **past loans** only. The approach was to treat them as new loans, by ”turning back the clock” on the data by excluding any information that was sourced after the loan was initiated

- Project goal:

  Provide a predictive model and actionable insights on potentially risky loans for LendingClub investors, at the time they actually need it – **before** they decide **on which loan to put the money on**

Data Exploration and Preliminary Analysis

Defining the target – what is a “Default loan”?

- The Dataset (161,231 unique loans) most are **Current** loans (~94%) whereas **Fully paid** are (4%) and the rest are in 5 more categories*

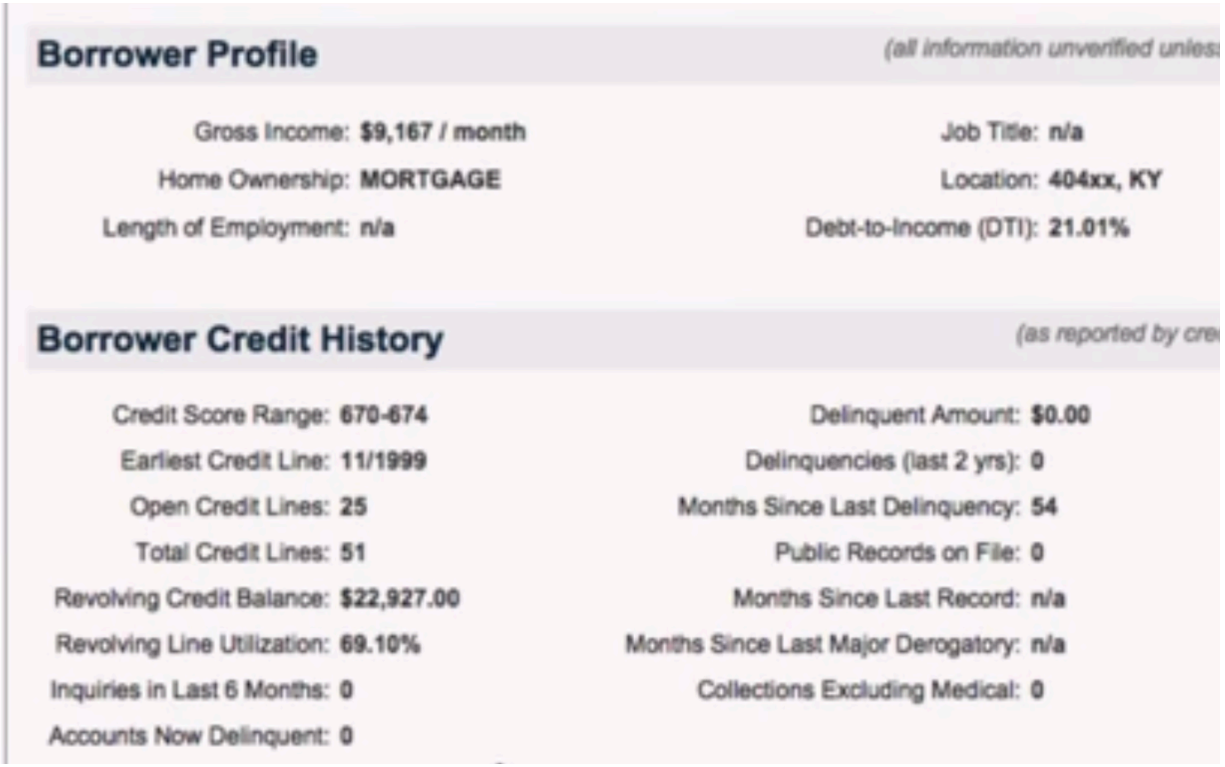
- A search on LC website** gave the basis for our following assumption:
  Since 75% of **late (31-120 days)** loans end up not fully paid, we aggregated them with **Default** group and **Charged-off** group to make up a diverse “bad loans” group we named “**Default”**

***Note:** Our assumption is that using only a subset of the available data serves best the business goal we chose

Data Exploration and Preliminary Analysis

Our feature definition – A variable that is a available for investors on a loan before it is issued

Right:
Fields available on LC website On a loan for a 14 day window, before it is issued (or discarded due to lack of demand)



Borrower Profile *(all information unverified unles*

Gross Income: **$9,167 / month**
Home Ownership: **MORTGAGE**
Length of Employment: **n/a**

Job Title: **n/a**
Location: **404xx, KY**
Debt-to-Income (DTI): **21.01%**

Borrower Credit History *(as reported by cre*

Credit Score Range: **670-674**
Earliest Credit Line: **11/1999**
Open Credit Lines: **25**
Total Credit Lines: **51**
Revolving Credit Balance: **$22,927.00**
Revolving Line Utilization: **69.10%**
Inquiries in Last 6 Months: **0**
Accounts Now Delinquent: **0**

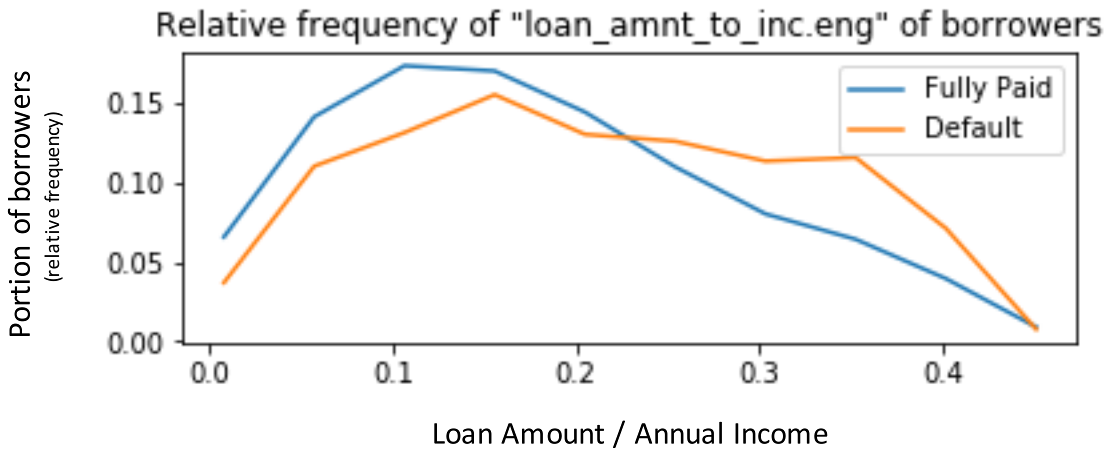
Delinquent Amount: **$0.00**
Delinquencies (last 2 yrs): **0**
Months Since Last Delinquency: **54**
Public Records on File: **0**
Months Since Last Record: **n/a**
Months Since Last Major Derogatory: **n/a**
Collections Excluding Medical: **0**

** https://www.lendingclub.com/info/demand-and-credit-profile.action

Data Exploration and Preliminary Analysis

Exploring dataset features

We hypothesized that borrowers with high **loan amount relative to their income** should have more difficulties in paying back their loan. We see evidence of that by exploring loan to income ratio:

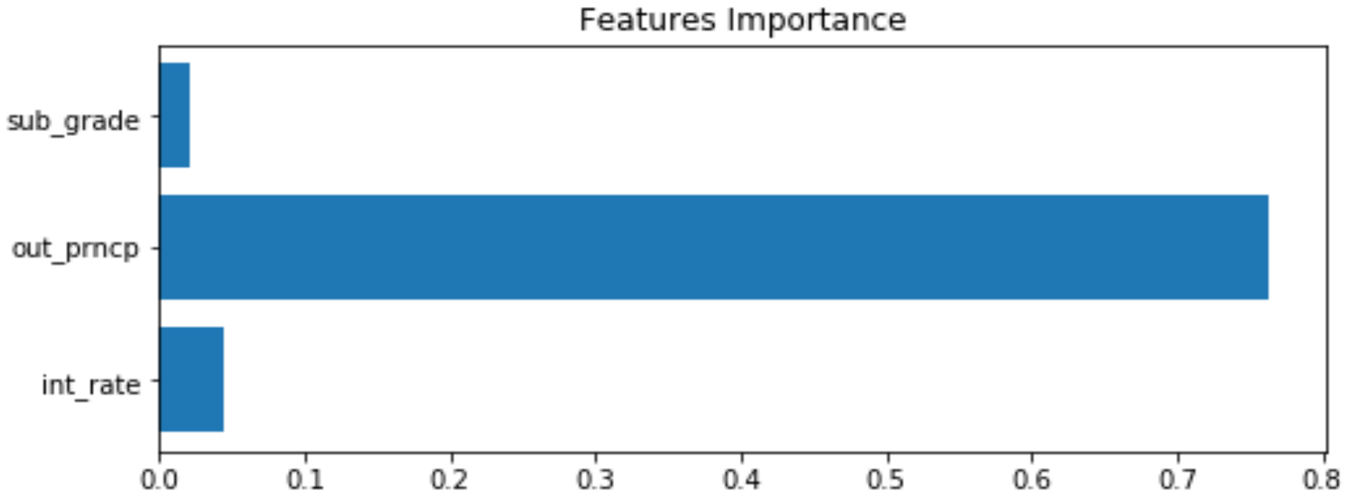


Fully Paid: μ=0.20, σ=0.1
Default: μ=0.24, σ=0.11

Data Exploration and Preliminary Analysis

Pinpoint features NOT to be used

- Target-leaked features tend to be suspiciously strongly correlated with the target*, we therefore need to detect them, as done here by running a naïve model on the dataset and extracting the features importance



'out_prncp' is the remainder of borrowed amount that is still needed to be paid. By running a naïve model on the dataset that includes this feature we see how 75% of the prediction relies on this feature. **Default** loans by definition still have to be payed. whereas **fully paid** loans by definition have nothing sum left to be payed.

* Target leak - means that when deploying our model for prediction we wouldn't have the feature available to us because it is determined together with the target (hence "target leak).

Baseline modeling

Classification with raw data

- Initial classification efforts with the raw dataset shows that the ability to detect Defaults (recall) is 0% (classifying every loan as fully paid)
- This means that without feature engineering and class-balancing efforts (# of loans in each class) we can’t get a sufficiently sensitive classification*

***Note:** we can use a target-leak feature in the dataset as a type of a “positive-control”. we’ve reached a 77% rate of detection with one such feature, showing a ballpark of expected detection capabilities.

Feature Engineering

Selected Engineered Features:

- log_ratio (annual inc / state median income)
- Log Credit years
- Months since last major derogatory
- Principal
- Home Ownership and Purpose
- Grade & sub grade

Feature Engineering

Selected Engineered Features:

- log_ratio (annual inc / state median income)
- Log Credit years
- Months since last major derogatory
- Principal
- Home Ownership and Purpose
- Grade & sub grade

How shall we use location data?

- Both zip code and state should be informative
- zip code has has 3 digits only (low resolution)
- Ideally, if this project should continue we would engineer:
  I. With longitude & latitude converter
  II. Enrich with census data (translate zip to socio-economic data into categorical features)
- State is less informative than zip, but we can quickly use a proxy such as median income per state (in 2014*) that may show the socio-economical status of the residents of this state
- We also adjusted the annual income of borrowers to the state median to input their relative income which might be more informative

* https://www.reinisfischer.com/median-household-income-us-state-2014

Feature Engineering

Selected Engineered Features:

- log_ratio (annual inc / state median income)
- Log Credit years
- Months since last major derogatory
- Principal
- Home Ownership and Purpose
- Grade & sub grade

How to add the credit longevity to the data?

- Both Date of issue of loan & earliest credit line date could not be used as is
- Subtracting earliest credit line date from issue date gives us the years a borrower has credit
- Conducted log transformation (we assume that difference between 5 -10 years is more significant than 15-20 years)

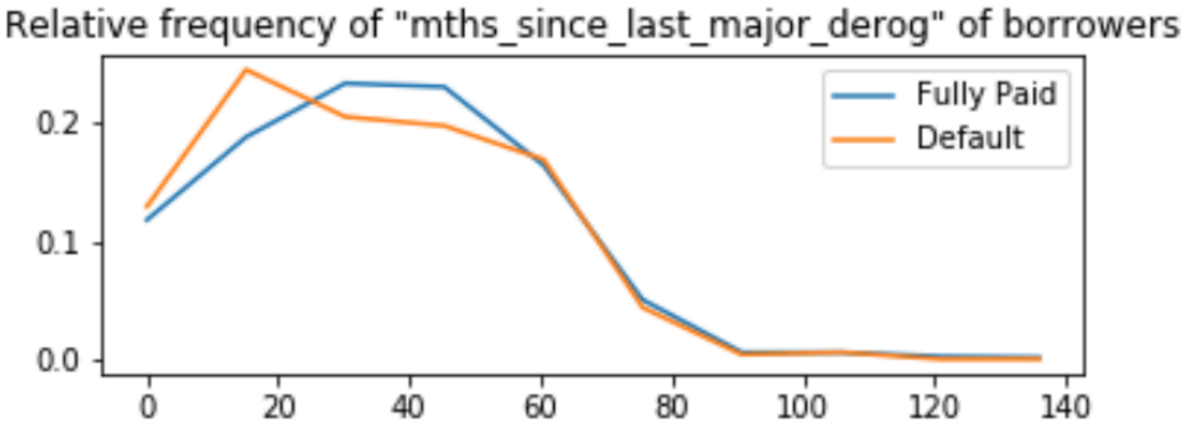
Feature Engineering

Selected Engineered Features:

- log_ratio (annual inc / state median income)
- Log Credit years
- Months since last major derogatory
- Principal
- Home Ownership and Purpose
- Grade & sub grade

Can a feature with mostly missing data be informative?

- the data that is not missing seems informative up to 60 months mark
- There is a intersection at 24 months mark where the relative frequency of a default loan diminishes further, while fully paid loans increases and continues to be high
- Categorical transformation to 'recent' (<25) and 'old' (24 → 60) while the rest was categorized in 'unknown' group together with missing instances



Feature Engineering

Selected Engineered Features:

- log_ratio (annual inc / state median income)
- Log Credit years
- Months since last major derogatory
- Principal
- Home Ownership and Purpose
- Grade & sub grade

Other selected feature transformations

- Principal was calculated by installment – (interest rate*) installment. The rational was to isolate the monthly payment of the loan from credit aspects (the int rate which is partly depended on credit rating

Feature Engineering

Selected Engineered Features:

- log_ratio (annual inc / state median income)
- Log Credit years
- Months since last major derogatory
- Principal
- Home Ownership and Purpose
- Grade & sub grade

Other selected feature transformations

- Home ownership and purpose categories were expanded to “dummy” features
- Grade & subgrade were transformed to numerical ranking

* **Note:** Additional features where transformed and engineered. The full list of features and their description is available in a separate document

Building a Predictive Model

- **Assumption** –
  - Lending Club investors would like to detect the highest number of risky loans possible, rather than avoid a false alarm (falsely identifying a default loan where in reality it is not)
  - This is because of the financial toll related to loans not payed back, as compared to the occasional investment miss-opportunity when mistakenly avoiding a loan because of false alarm

Building a Predictive Model

- **Assumption** –
  - Lending Club investors would like to detect the highest number of risky loans possible, rather than avoid a false alarm (falsely identifying a default loan where in reality it is not)
  - This is because of the financial toll related to loans not payed back, as compared to the occasional investment miss-opportunity when mistakenly avoiding a loan because of false alarm

- **Method** –
  - We ran several classification models (i.e. Logistic Regression, Random Forest (RF)) on the dataset to get initial idea about the predictive power of the models
  - Although RF showed inferior baseline performance as compared to logistic regression, RF is generally a more robust model*
  - We chose AUCROC as a metric for performance due to its robustness**
  - Then we optimized the model depended on our metric (AUCROC score)
  - Finally we'll set our recommendations based on several sensitivity boundaries

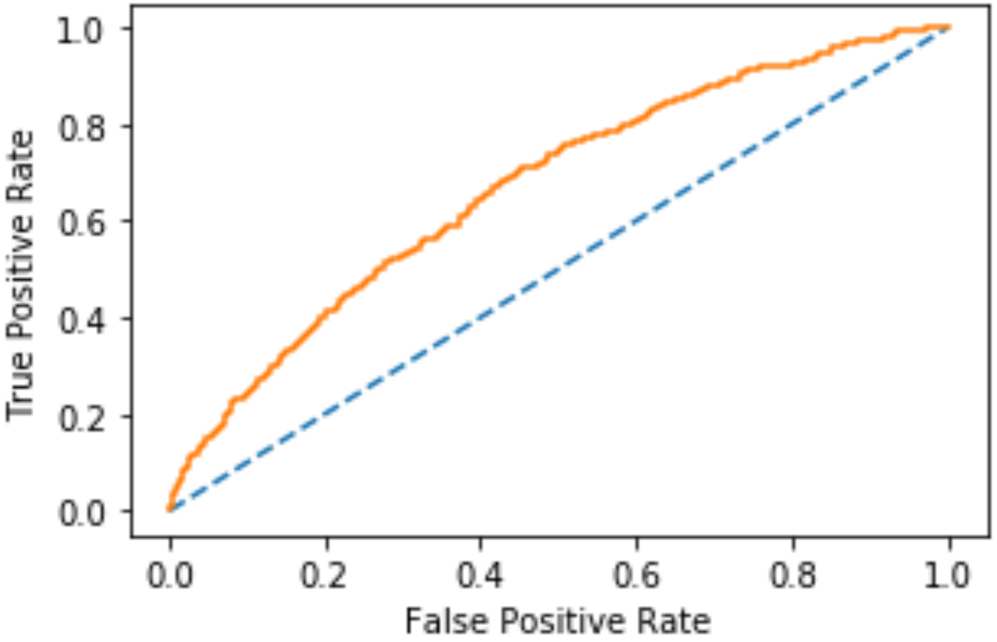
* Has more hyper parameters to tune so there is a better potential to improve its performance further. And can better cope with categorical features
** allows us to tune for the best model across all sensitivities and precisions

Model Performance

- **What is the performance of the model we are offering?**

ROC curve in orange shows the range of different sensitivities we could choose from. Since we are interested in high sensitivity, we can choose a a high FP rate (e.g. around 0.6 will give us ~80% sensitivity).

	Mean	STD
CV AUC score	0.65	0.15



Summary and Final Recommendations

- **What the model is revealing to us about Default factors?**

1. Interestingly renting and owning are related to risk while **mortgage to de-risk** (counter to our hypothesis), perhaps this group have higher credit scores (which allowed them to get a mortgage in the first place)
2. **Income verification** is related to default (counter intuitive) perhaps the verification is done only when there is a raised suspicion which is understandably related to default
3. Negative financial history in the last 2 years is related to default whereas between 2 and 4 years is related to **paid loans**

* **Note: It's** Questionable whether this is an actionable insight. it's possible that whole-loans are grabbed "in one go" because they are more appealing. So in P2P it's probably easier to identify whether a loan is fractional (cause it's too late or too early to be whole)

Factor	Related to Default
Low credit score / high interest / fractional loans *	+++
Income verified or source verified [2]	++
Recent [3] (or unknown): delinquency, major derogatory or record	+
Renting, owning home [1] / debt consolidation / low annual income	+
Years of employment / whole loans	- - -
Mortgage	- -
Older: delinquency, major derogatory or record [3] / Term 60	-

Refer to supplementary slides for a background on this table

Summary and Final Recommendations

- **Final conclusions and actionable recommendations for LendingClub investors**

- The sensitivity of the model we are offering to LendingClub investors ranges from identifying 60% to 80% of default loans
- The desirable default identification rate is depended on the preferences of investors whether they are risk averse or risk takers
- We offer 2 possible packages:

  i. For risk-averse investors we offer the 80% default detection package. While this means that more than half (60%) of the loans that will be labeled by us as risky will end up to be perfectly safe loans, investors taking this package would enjoy lowering their risk to just 1.2% (from 6% according to the rate of default loans on Lending Club)

  ii. Alternatively, for risk taking investors we offer the 60% default detection package. Which means that 40% of the loans that will be labeled by us as risky will end up as safe, but investors taking this package would enjoy lowering their risk by almost half - from 6% to 3.6%

Summary and Final Recommendations

- **Final conclusions and actionable recommendations for LendingClub investors**

- If investors would like to pick lower-risk loans manually (without using our packages) but still would like to benefit from our analysis, We recommend to follow these guidelines:

I. Take into consideration low credit cores first

II. whether the income or the income source **are** verified (as they are more related to default than not)

III. Lookout for negative credit history in the last 24 months. If it’s later than 24 months it’s **even safer** than loans with unknown history.

IV. Counterintuitively owners of homes have **higher risk** than mortgage payers

Next Steps

- Continue Feature engineering efforts
- The data could be enriched further with publically available data, e.g. Implementing data on locations using zip etc.
- Run the model on the ”current” data → explore the features that are related to default class → try to find new features that are correlated to the default class between
- Try different classification models, Do residual analysis on current model
- “Anomaly detection” on current data

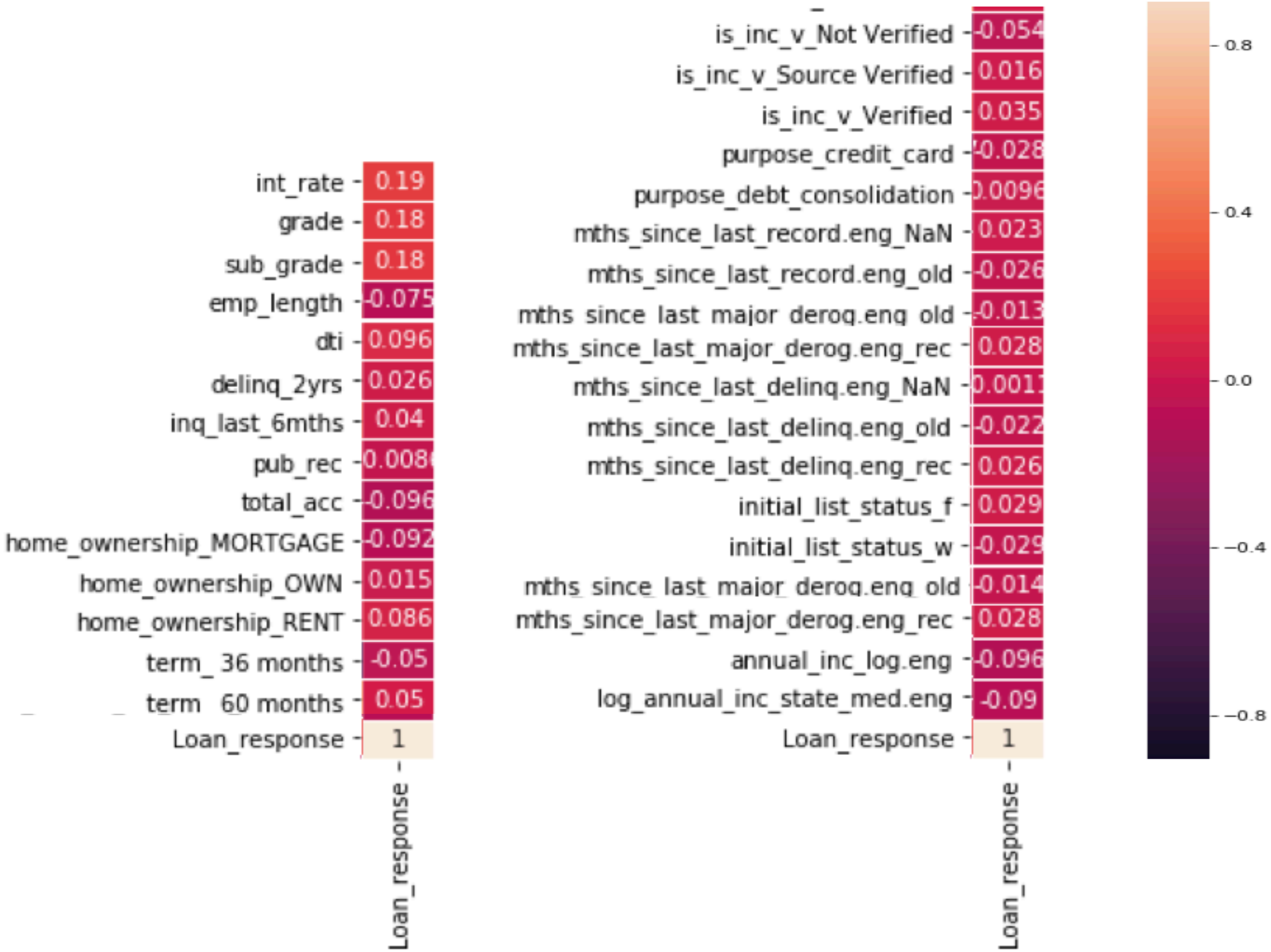
Q & A

Supplementary slides

Summary of the Proposed Model and Final Recommendations

- **What the model is revealing to us about Default factors?**

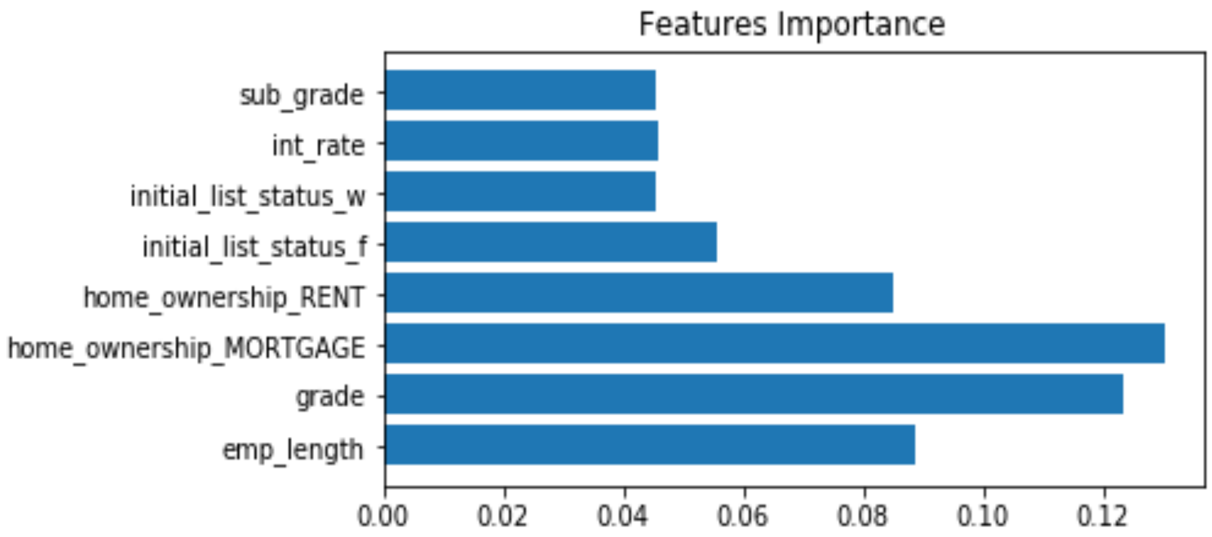
A heat-map of correlations can give a better sense of how these factors are linked to defaults



Summary of the Proposed Model and Final Recommendations

- **What the model is revealing to us about Default factors?**

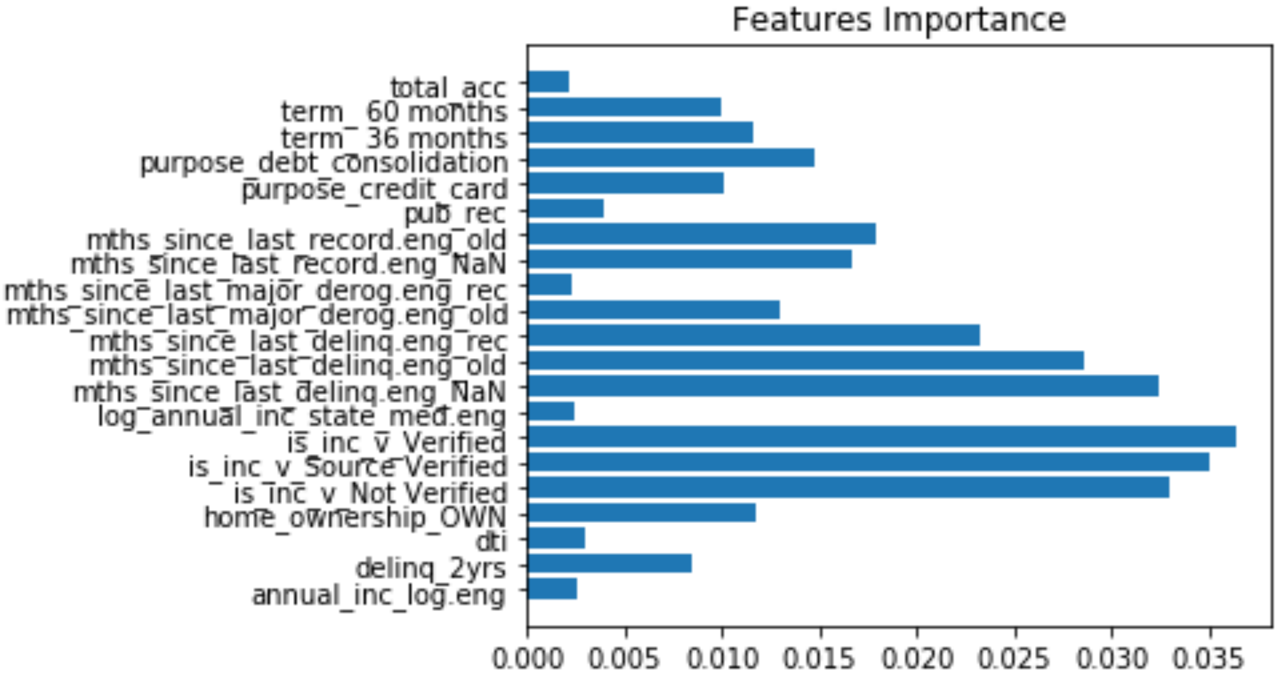
- **Credit score related features** (grade, subgrade and int_rate) play an important role in the model, affirming that on average, credit score counts as expected (although it might put borrowers at risk in the first place due to higher monthly fees).
- Even a stronger predictor is **homeownership status**, we suggest that the reason is that it defines more strictly the borrowers (renters and mortgage payers have high monthly payments)



Summary of the Proposed Model and Final Recommendations

▪ What the model is revealing to us about Default factors?

- Whether **income is verified is** important
- **Negative credit history related features** (e.g. months since last 'delinq', 'last record' and 'major derog') are also important Even though all of them suffered from significant missing data
- A longer **term** might mean that borrowers with longer terms are the ones with difficulty of returning
- Surprisingly log of annual income and dti are not as "important" as hypothesized (and adjusting to median income of state might not add information)
- Also surprisingly absent are the number of years of credit, and balance utility



Prediction ability isn't necessarily depended on few important features, rather on multitude of features that each might play a small part not be